JASA ARTICLE

Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing

John H. L. Hansen,[a] Marigona Bokshi,[b] and Soheil Khorram[b]
Robust Speech Technologies Laboratory (RSTL), Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, Texas 75080, USA

ABSTRACT:
Speech production variability introduces significant challenges for existing speech technologies such as speaker identification (SID), speaker diarization, speech recognition, and language identification (ID). There has been limited research analyzing changes in acoustic characteristics for speech produced by untrained singing versus speaking. To better understand changes in speech production of the untrained singing voice, this study presents the first cross-language comparison between normal speaking and untrained karaoke singing of the same text content. Previous studies comparing professional singing versus speaking have shown deviations in both prosodic and spectral features. Some investigations also considered assigning the intrinsic activity of the singing. Motivated by these studies, a series of experiments to investigate both prosodic and spectral variations of untrained karaoke singers for three languages, American English, Hindi, and Farsi, are considered. A comprehensive comparison on common prosodic features, including phoneme duration, mean fundamental frequency (*F*0), and formant center frequencies of vowels was performed. Collective changes in the corresponding overall acoustic spaces based on the Kullback-Leibler distance using Gaussian probability distribution models trained on spectral features were analyzed. Finally, these models were used in a Gausian mixture model with universal background model SID evaluation to quantify speaker changes between speaking and singing when the audio text content is the same. The experiments showed that many acoustic characteristics of untrained singing are considerably different from speaking when the text content is the same. It is suggested that these results would help advance automatic speech production normalization/compensation to improve performance of speech processing applications (e.g., speaker ID, speech recognition, and language ID). © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0001526

(Received 25 October 2019; revised 17 June 2020; accepted 19 June 2020; published online 18 August 2020)
[Editor: Paavo Alku] Pages: 829–844

I. INTRODUCTION

Research has shown that speech production variability due to stress, speaking style, emotion, Lombard effect (speech produced in noise), cognitive/physical conditions, etc. all impact speech technologies for speech recognition, speaker identification (ID), etc. (Hansen and Hasan, 2015; Hansen, 1996). For speech processing, studies have explored ways to compensate for speech under stress production variabilities, including cepstral feature compensation (Hansen, 1996; Bou-Ghazale and Hansen, 2000), training token generation (Hansen and Bou-Ghazale, 1995), and combined stress compensation and classification (Womack and Hansen, 1999). These methods have helped address speech production variability for speech technologies but often require parallel datasets to achieve effective compensation. Neural network based solutions were also considered in the past (Hansen, 1996). An area that also saw progress in analysis of production variability is automatic stress detection, generally, in a text and speaker independent framework. One of the earliest studies on the analysis of production variability was based on the nonlinear Teager energy operator (TEO; Cairns and Hansen, 1994; Zhou *et al.*, 2001), which was later expanded to incorporate weighted frequency sub-bands (Hansen *et al.*, 2011), and then used hybrid tracking and classification schemes (Hansen *et al.*, 2012). These approaches established viable speech processing methods to explore speech production variabilities, including vocal effort such as a whisper (Ghaffarzadegan *et al.*, 2016), physical task stress (Godin and Hansen, 2011), and airflow dynamics based on a physiological microphone (PMIC; Patil and Hansen, 2010). These studies have collectively explored a range of speech production changes which are generally intrinsic in nature (Hansen and Hasan, 2015). One research question that arises is if a speaker intentionally alters his or her speech production mode, does this impact the performance of general speaker models for speaker recognition? To address this question, this study considers one form of mismatch between neutral speech production and an altered style. Specifically, we seek to address the question of what changes in the speech production if a speaker reads text content versus sings the same text sequence? As such, the main research goal here is to quantify the differences between speaking and singing the same text content and how this impacts speaker recognition systems using neutral trained speaker models.

[a]Also at: Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA. Electronic mail: john.hansen@utdallas.edu, ORCID: 0000-0003-1382-9929.
[b]Also at: Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA.

In recent years, there has been an increased interest in speech production variability, including the analysis of speech production differences between speaking and singing (Peynircioğlu *et al*., 2017; Beeman, 2017; Mehrabani and Hansen, 2013a). These studies have shown considerable deviation in speech production based on spectral and prosodic characteristics of professional singing from speaking (Carlsson and Sundberg, 1992; Story, 2004). The main difference has been found in the acoustic characteristics of vowels (e.g., intensity and fundamental frequency of vowels; Fowler and Brown, 1997; Bloothooft and Plomp, 1984). Sundberg *et al*. have also reported a major difference in formant frequencies of singing versus speaking (Sundberg, 1977; Sundberg, 1974). In addition, researchers have developed various engineering applications in areas such as automatic singer ID and music language recognition in order to address the problem of automatic music labeling and retrieval (Tsai and Lee, 2012; Tsai and Wang, 2004; Mehrabani and Hansen, 2011, 2013b; Sangwan *et al*., 2011).

Previous papers have explored the challenge of identifying the production differences between singing and speaking and, although very comprehensive, considered only trained singers (Watts *et al*., 2006; Bloothooft and Plomp, 1986; Carlsson and Sundberg, 1992; Saitou *et al*., 2007). Although these studies hold great scientific value, they focus on pinpointing specific factors relevant to trained singers and lack the focus needed to explore singing as a variation of neutral speech production in general (Sundberg, 1977; Sundberg, 1974). Little to no effort has been dedicated in considering the singing of untrained singers and how it deviates from typical neutral or conversational speech. The results of analyzing the untrained singing voice can be used to improve various speech processing applications, such as automatic speech recognition and speaker identification (SID), since it represents a significant production mismatch from neutral speech. It also holds scientific and forensic benefits because it would provide further knowledge to formulate speaker embedding models for voice forensics, as well as being able to link versions of singing voices to speaker voices in music authentication. Generally, the performance of speech processing systems normally degrades significantly when introduced with speech types that highly deviate from regular/neutral speech. Extensive studies have been performed for speech production under stress, Lombard effect (Hansen, 1996; Hansen and Hasan, 2015; Cairns and Hansen, 1994; Zhou *et al*., 2001; Hansen *et al*., 2011), emotions (Khorram *et al*., 2016; Khorram *et al*., 2018), etc. These studies introduce compensation strategies to improve automatic speech technology under mismatched conditions (Hansen, 1996; Hansen and Hasan, 2015).

A number of studies have focused on the voice characteristics that help singers to be heard more distinctly in large venues, such as opera houses, even over a high level of sound from the orchestra (Sundberg, 1977; Barrichelo *et al*., 2001; Cleveland and Sundberg, 1999). In particular, Sundberg (1977) provides a comprehensive study to identify which aspects of the singers' voice are unique, showing that there is a major difference between the way vowels are pronounced in singing and the way they are pronounced in spoken speech. Barrichelo *et al*. (2001) explored if trained singers can carry their formant ability into their conversational speech. This same study then explored the differences between the third and fourth formants of trained singers and untrained speakers.

Another important limitation of previous studies is that they used data from only one specific language (Barrichelo *et al*., 2001); therefore, their results were valid for that specific language, which limits their impact since they cannot analyze cross-language dependencies. Leveraging a multilingual database helps us to explore the differences between singing and speaking in a broader scope and allows for the analysis of cross-language dependencies among untrained singers.

The main goal of this paper is to quantify the differences between speaking and singing signals which impact SID and, potentially, language ID systems. In this study, we considered the production differences between speaking and untrained karaoke-style singing by collecting a multilingual database with speakers in three languages: American English, Hindi, and Farsi. We considered singing as a special category of neutral speech with the goal of providing baseline knowledge that is useful in improving the performance of speech recognition, SID, and language ID systems. We analyzed the prosodic components that suffer the greatest change when individuals migrate from a speaking style to singing: temporal duration, formant frequencies, and fundamental frequency (pitch). We also employed the Kullback-Leibler (KL) divergence (Hershey and Olsen, 2007) between multidimensional probability distribution models trained for singing and speaking to quantify the differences in their practical production spaces. The same type of model was also used for a GMM-UBM (Gausian mixture model with universal background model) SID task (Hansen and Hasan, 2015; Reynolds, 1995). By collecting a multilingual corpus, it will be possible to explore the cross-language differences between the speakers. One potential application of this study would be the formation of an automatic singing skills assessment strategy of untrained singers that aims to determine a speaker's future singing potential.

Many speech processing engines are helping us to do our work better and have a better life. They are ready to play their important role as a critical part of human computer interaction systems by listening to us all day. We do not necessarily communicate with the speech processing engines through read speech, and our speech may take different forms. For example, our speech may be emotional, whispered, and in a singing form. Most of the people have not been trained for singing, and it is important to be able to analyze their singing speech. At least, it is essential to know the performance of the current speech processing systems when they are processing singing signals of unprofessional singers. However, there has not been any research on analyzing voice characteristics of the untrained singers, and there is a need for that. For example, many voice activated devices or dialogue systems have been equipped with a

speaker recognition module, which helps them to make better decisions by considering speaker-specific characteristics. These speaker recognition modules may need to process singing voices of untrained singing signals; however, they have not been developed and tested for that. It is of great importance to evaluate and improve the performance of the speech processing tools for the untrained singing signals, and this paper is the first step toward this important goal.

II. BACKGROUND

Most research studies in singing consider the production space of vowels as they considerably vary during singing (Sundberg, 1977; Story, 2004). For example, Mehrabani and Hansen (2013a) studied how the intrinsic dimensionality of vowels changes from speaking to singing. That study applied locality preserving projection (LPP) for dimensionality reduction of the spectral feature vectors and conducted a vowel classification experiment in the low-dimensional subspaces. The results of the classification experiment showed that a higher number of intrinsic human speech/voice production dimensions is required for effective representation of singing vowels.

A study by Brown *et al.* (2000) investigated how well a group of listeners can differentiate between the spoken and sung segments of professionally trained singers and non-singers. They used a comprehensive set of acoustic measures in their study, including fundamental frequency, duration, percent jitter, percent shimmer, noise-to-harmonic ratio, the singer's vibrato, and the singer's formant. The results show that professional singers are easily distinguishable from the non-singers by their singing voices but poorly identifiable by their speaking voices.

An important characteristic of the professional singers' voices is defined by the term "*singer's formant*." According to Barrichelo *et al.* (2001), this formant occurs when the third, fourth, and fifth formants are close in frequency and we observe a peak in the frequency of 3 kHz of vowels; Fant (2012) also found that the third and fourth formants are closer in trained singers' voices. In order to produce these formant frequencies, the physical configuration of the vocal tract must be changed. The singer's formant is an important feature that helps classically trained singers to be heard over a full orchestra.

Not all professional singers have the ability to generate the singer's formant. A study in Cleveland and Sundberg (1999) explored the relationship between the first four formants in singing and speaking among the singers of country music. The results showed that for most well-renowned country singers, the correlation between the first four formant frequencies in singing and speaking was high, indicating that, for the most part, country singers use similar formant frequencies in speech and singing. If there existed a difference, the formant frequencies in singing would be higher than those in speaking, and this was attributed to the fact that the fundamental frequency for these singers was higher.

Sundberg (1977) introduced a number of factors that make the singer's voice special. They explained the radiated sound of singing through the properties of the voice-source spectrum and the formants of singing. Sundberg (1977) showed that the major difference happens in the way the formant frequencies are produced in speech versus singing and the way that vowels are pronounced in both cases. More precisely, they found that the first two formant frequencies are normally lower in the singing version of a word and the spectral energy is considerably higher between 2.5 and 3 kHz.

The question that most research in the engineering area attempts to address is: given a set of training tokens with neutral speech from different singers, how well can an automated SID system determine the identity of the singer when provided with a singing test segment? We know, in reality, that given a segment of a famous singer's voice, one can determine the identity of the singer fairly quickly if we are familiar with the singer. We can probably even determine his or her identity if we hear him or her speaking. This criterion has driven most of the work in automated singer ID and is a motivation for us to include some results in this area as well.

Considerable investigations in SID for singers has also been accomplished in a number of studies (Zhang, 2003; Tsai and Lee, 2012; Tsai and Wang, 2007). The study in Zhang (2003) is one of the first works in this area to consider the identity of a singer by analyzing the audio features of the music signal. This work includes the application of digital music databases and retrieval and provides powerful functions for use in browsing and searching musical content. By providing the identity of a speaker in a music database, one can then retrieve all songs sung by a particular singer. There are some questions that could arise, such as if there are multiple singers/groups, songs are solos, etc. The same technology is used to cluster songs of similar voices of singers or search for songs which are similar to a query song in terms of a singer's voice (Zhang, 2003). The approach taken in that study is based on the fact that the time-frequency features of a singing voice are quite different from those of a speaking voice even when the actual text context is the same. Therefore, the concept is that by extracting and analyzing audio features properly, an automatic system should be able to achieve a certain degree of singer ID as well. The features used in this system consisted of Mel frequency cepstral coefficients (MFCC), which are the most commonly used features in SID and speech recognition systems due to their ability to effectively represent the configuration of the vocal tract of a speaker (Zhang, 2003).

The work by Tsai and Lee (2012) is in the area of singer ID based on speech derived models. The study investigated the possibility of modeling a singer's voice using spoken data instead of singing data. The motivation for this work was the result of the limitation in available singing data of trained singers compared to simply collecting spoken samples of these speakers. The difficulty in obtaining this model is due to the fact that singing by a person deviates

significantly from a typical spoken speech style. Therefore, in order to address this issue, a system was proposed that uses maximum *a posteriori* (MAP) adapted models based on a limited set of singing data. The results in this aforementioned study showed a significant increase in performance when presented with this alternate classification technique.

III. DATASET

In the current study, we used the University of Texas (UT)-Sing database collected at the Center for Robust Speech Systems (CRSS), University of Texas at Dallas. The database contains more than 23 hours of speech from 81 speakers across 4 different languages: English, Farsi, Hindi, and Mandarin. As noted, the dataset contains material from the Mandarin language, but that portion of the dataset had not been transcribed yet (i.e., phoneme level transcription is needed) and, therefore, we did not use that portion in the current experiments. We asked the speakers to choose from a list of five popular songs, which they were asked to read/speak the words through a karaoke display system in their own language (e.g., no background music played through open-air headphones), followed by the same song, but this time they were asked to sing with background music played through the karaoke system (again, using open-air headphones so all speaker content is music-free). The speakers were not professional singers, but they were completely familiar with the selected songs they chose. Again, speakers had access to the lyrics of the same song through a prompt display for both reading and singing. We selected 16 songs for each language, on average, with some overlap between the intra-language songs. This setup allowed us to make a fair comparison between the speaking and singing production differences of each speaker.

Since male and female speakers have, generally, different speech production systems, we separated the data based on gender and ensured that the speaker labels reflected this separation. In order to have natural karaoke singing, we played background music of each song for the speakers through open-air headphones. Since the open-air headphones were employed, there was a negligible occlusion effect. However, the open-air headphones may result in leakage on the played back music and, therefore, we accurately investigated the amount of music leakage in the recorded audio files. We noticed that the amount of this leakage is not significant as the signal to music ratio of the audio files is not less than 10 dB. We believe this high signal to music ratio is a result of the high-quality microphones (Shure Beta-54 close-talk microphone, Niles, IL) that we have used during our data collection process. Karaoke system prompts were used during this process so that as the speaker would sing or read, the lyrics of the song were displayed on the screen.

The acoustic data were recorded in a soundbooth using a Shure Beta-54 close-talk microphone (Niles, IL; Hansen and Varadarajan, 2009). The Shure Beta-54 is not completely compatible with the recommendations explained in Patel *et al.* (2018) and Svec and Granqvist (2010)], but for the type of analysis that we perform in this paper, e.g., SID analysis, there is no need to follow the recommendations of Patel *et al.* (2018) and Svec and Granqvist (2010). The Shure Beta-54 microphones are highly directional (super-cardioid), and it is important for us to have omnidirectional microphones to prevent the music leakage from our open-air headphones while recording the sound of our speakers. We asked the speakers to confirm that they had never taken any voice lessons or vocal training of any sort before participating in the study. The speakers were then asked to select five songs that they were familiar with from the list of popular songs provided. This allowed us to focus our analysis only on the singing versus speaking production differences without capturing other factors, such as accent, intonation, speaker familiarity with song, etc.

In this study, in order to explore the specific phonemic language dependencies across speakers while singing and speaking, we used speakers from three of the languages contained in our database: English, Hindi, and Farsi. Here, we selected six speakers from each language (three males and three females). We used phonetically transcribed data of all speakers to conduct our experiments based on (i) phoneme duration, (ii) fundamental frequency, and (iii) formant frequency features.

All sung and spoken utterances of each speaker were manually annotated by three fluent transcribers in English, Hindi, and Farsi. The annotations were performed using Wavesurfer (Sjölander and Beskow, 2000), a speech analysis tool equipped with transcription labels and spectral sessions in the form of spectrograms that facilitate the determination of phoneme boundaries. For vowel analysis, we used the most dominant vowels in each of the three languages. Since our analysis is based on sustained vowels, we only considered the vowels with a duration of more than 0.1 s. For reliable cross-language results, we picked vowels from approximately the same production space area in each language.

As this paper presents the first study on the UT-Sing dataset, we explain more statistics and characteristics of the UT-Sing dataset in this section. Data from three different languages have been used in the current study: English, Hindi, and Farsi. For each language, we recorded data from six different speakers. In terms of gender, the dataset is balanced and, therefore, we used data from three males and three females for each language. All speakers were between 20 and 35 years old and they were all paid. For each speaker, we selected five songs, and all the speakers were completely familiar with the selected songs.

In order to record sounds, we used a Shure Beta-54 close-talk microphone (Niles, IL). All sounds have been recorded in 48 kHz sampling rate in mono 16-bit wave files; however, we converted all the sampling rates into 16 kHz before performing any analysis on the data. In addition, all the files have been recorded in a soundbooth and, thus, the quality of the recordings is high. According to our calculation, the signal to noise ratio of the files is between 10 and 40 dB. To calculate this number, we find the noise power from silent regions of our recordings and assumed all other parts have the same noise power. Next, we calculate the signal power by subtracting the noisy power from the noise power.

https://doi.org/10.1121/10.0001526

IV. PROSODIC ANALYSIS

According to previous studies, prosodic features of the singing voice are significantly different from prosodic features of neutral spoken speech (Sundberg, 1977). To explore this for untrained singing, this section presents a comprehensive study on various prosodic features of our recordings. In Fig. 1, we show the time-frequency changes of a randomly selected speaker's voice during singing and speaking. When

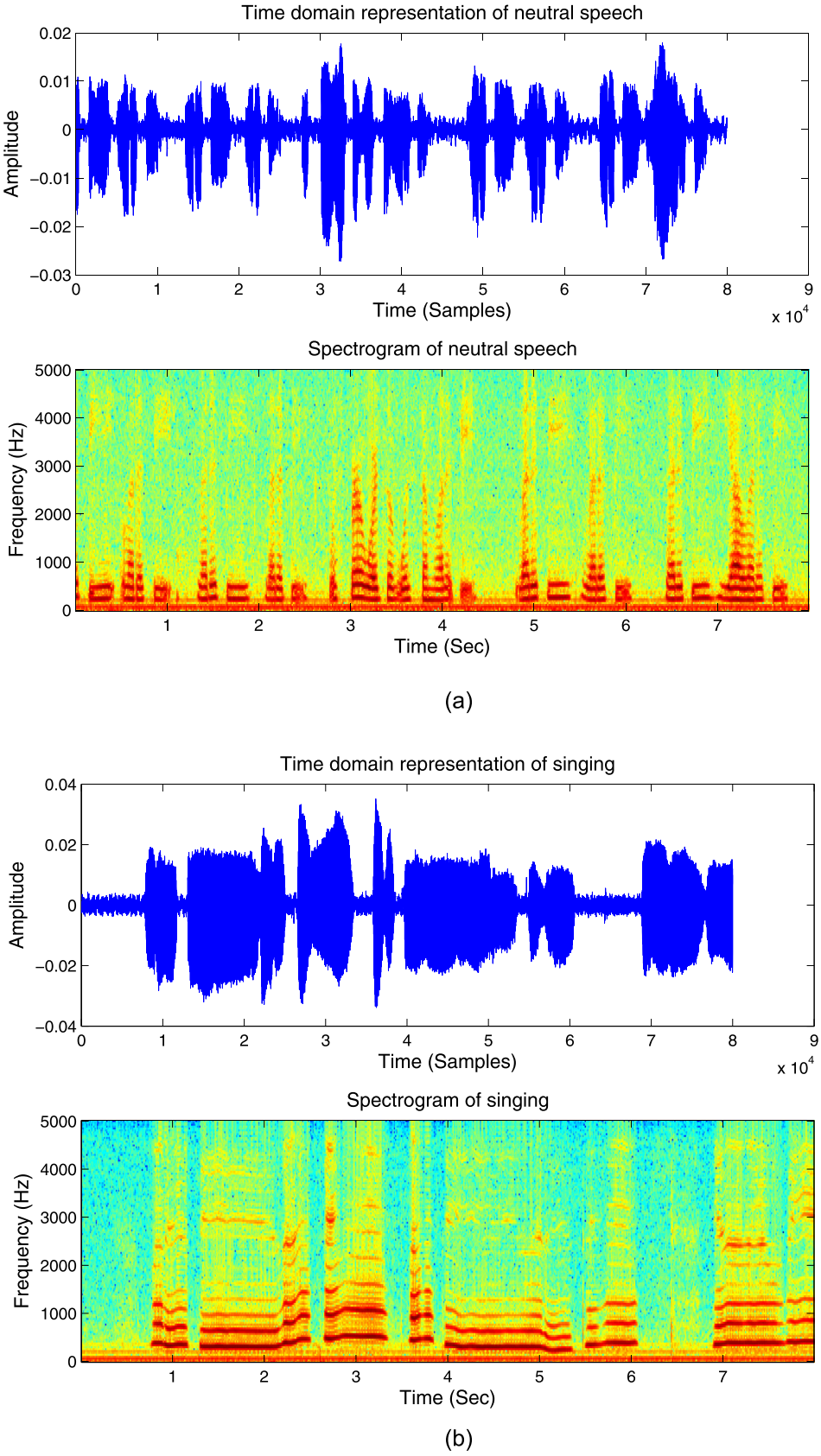


FIG. 1. (Color online) Time and spectrogram plots of the song excerpt "Let it be, let it be…," produced while (a) speaking and (b) singing. To plot this figure, a singing audio file was downsampled to 10 kHz and both time and spectral representations of the file were plotted. Therefore, every 10 000 samples of the time domain will be identical to 1 s. (a) Speaking "Let it be, let it be…," and (b) singing "Let it be, let it be…."

a person sings, their mean word duration and fundamental frequency increase. It has been shown that during singing the voiced to unvoiced ratio suffers the most in terms of change, and the fundamental frequency is used to represent a completely different state than in neutral spoken speech (Loscos *et al.*, 1999). In this section, we consider the temporal attributes and fundamental frequency in these two types of speech and compare their results across languages.

A. Temporal analysis

An important difference in the production space of singing is based on the temporal structure. Word duration increases during singing, and most of this increase is attributed to the elongation of vowels. It has been shown that the percentage of phonation time can increase from about 60% in speaking to 95% in singing (Loscos *et al.*, 1999). In our study, we showed the percentage change from speaking to singing for each phonetic class in English, Hindi, and Farsi. In order to achieve reliable results, we categorized phonemes into different classes and analyzed each class separately. For English and Hindi, the phonemes were grouped into eight phonetic classes: affricates, diphthongs, fricatives, stops, liquids, nasals, semivowels, and vowels. In Farsi, all phoneme classes do not exist (Khorram *et al.*, 2014; Khorram *et al.*, 2015), so we grouped phonemes into only five classes: affricates, fricatives, stops, nasals, and vowels. In English, there are 25 consonants and 20 vowels. Among these 25 consonants, 3 are nasals, 6 are stops, 2 are affricates, 9 are fricatives, and the remainder are laterals and approximants. In Hindi, there are 8 vowels, 2 diphthongs, 4 semivowels, 5 nasals, 4 fricatives, and 20 stops. In Farsi, there are six vowels, three nasals, nine stops, two affricates, nine fricatives, and four other consonants.

Table I summarizes the number of phoneme occurrences of each phonetic class in our dataset. In Table I, the count refers to the number of unique phonemes produced while speaking. Since the speakers sung and read the same lyrics, we assumed that the count for singing is similar (i.e., we assure no insertion/deletion of phonemes). As can be seen from Table I, vowels have the largest number of phonemes across all three languages.

An important factor that changes from speaking to singing is the phoneme duration. From the sing/speak duration ratio, it was realized that vowels show one of the largest increases in duration (3.0 for English, 2.7 for Hindi, and 2.8 for Farsi), followed by diphthongs, which have the largest duration ratio (3.2 for English and 3.3 for Hindi). In general, the average word duration increased by a factor of 2.0 from speaking to singing.

We also studied the relative duration of each phonetic class. If we consider an ideal word that contains a single entry from all eight phoneme classes (for English and Hindi) and five classes for Farsi, we can obtain the change in phonation time from speaking to singing for each phonetic class. In Fig. 2 and Fig. 3, we showed the changes in the percentage of the word duration for each phonetic group from speaking to singing for English, Hindi, and Farsi speakers.

As can be seen in Fig. 2, vowels and diphthongs occupy the largest percentage of the word duration in both English and Hindi for both speaking and singing. However, in Farsi, speaking affricates and fricatives have the largest percentage in word duration; this result changes for singing because vowels have the largest percentage of word duration in Farsi singing. The increase in duration for fricatives and affricates in the overall percentage of word duration during speaking in Farsi could be due to the amount of affricates and fricatives that are actually found in the Farsi language compared to vowels (e.g., in conversational Farsi, many written short vowels are not vocalized). However, during singing, vowels still occupy the largest percentage of word duration.

In English and Hindi, the only two phoneme classes that show a significant increase in duration are vowels and diphthongs. In Farsi, vowels show the largest increase and fricatives show a small increase of 0.3%. Since the increase in fricatives is too small to be statistically significant, we will say that only vowels show the highest statistically significant increase from speaking to singing. All other phonetic classes in all three languages show a decrease in percentage (although their actual absolute values do increase). The fact that vowels show the highest increase in all three languages explains the fact that singers tend to elongate the vowels when singing and, as a result, the other phonetic classes suffer a relative percent decrease in duration. In speaking, this elongation is not very typical and. consequently. the percentage in word duration is more evenly distributed across other phonetic classes.

The results of this experiment are important because they show the specific changes in the duration of each phonetic group, as well as how these changes differ across different languages. An important application of the results obtained in this section is that they can assist in spoken to

TABLE I. Phoneme count and phoneme duration ratio of singing versus speaking for English, Hindi, and Farsi. We used the value zero when the language did not have the particular phoneme class.

Phoneme	Affricates	Diphthongs	Fricatives	Liquids	Nasals	Semivowels	Stops	Vowels
English phoneme count	255	2306	3124	1795	2656	1281	3569	7160
English duration ratio	1.4	3.2	1.4	1.7	1.6	1.7	1.4	3.0
Hindi phoneme count	133	1534	1370	1346	1396	1600	3088	7142
Hindi duration ratio	1.3	3.3	1.2	1.6	1.6	1.6	1.2	2.7
Farsi phoneme count	129	0	2134	0	1560	0	2810	5545
Farsi duration ratio	1.2	0	1.4	0	1.7	0	1.4	2.8

https://doi.org/10.1121/10.0001526

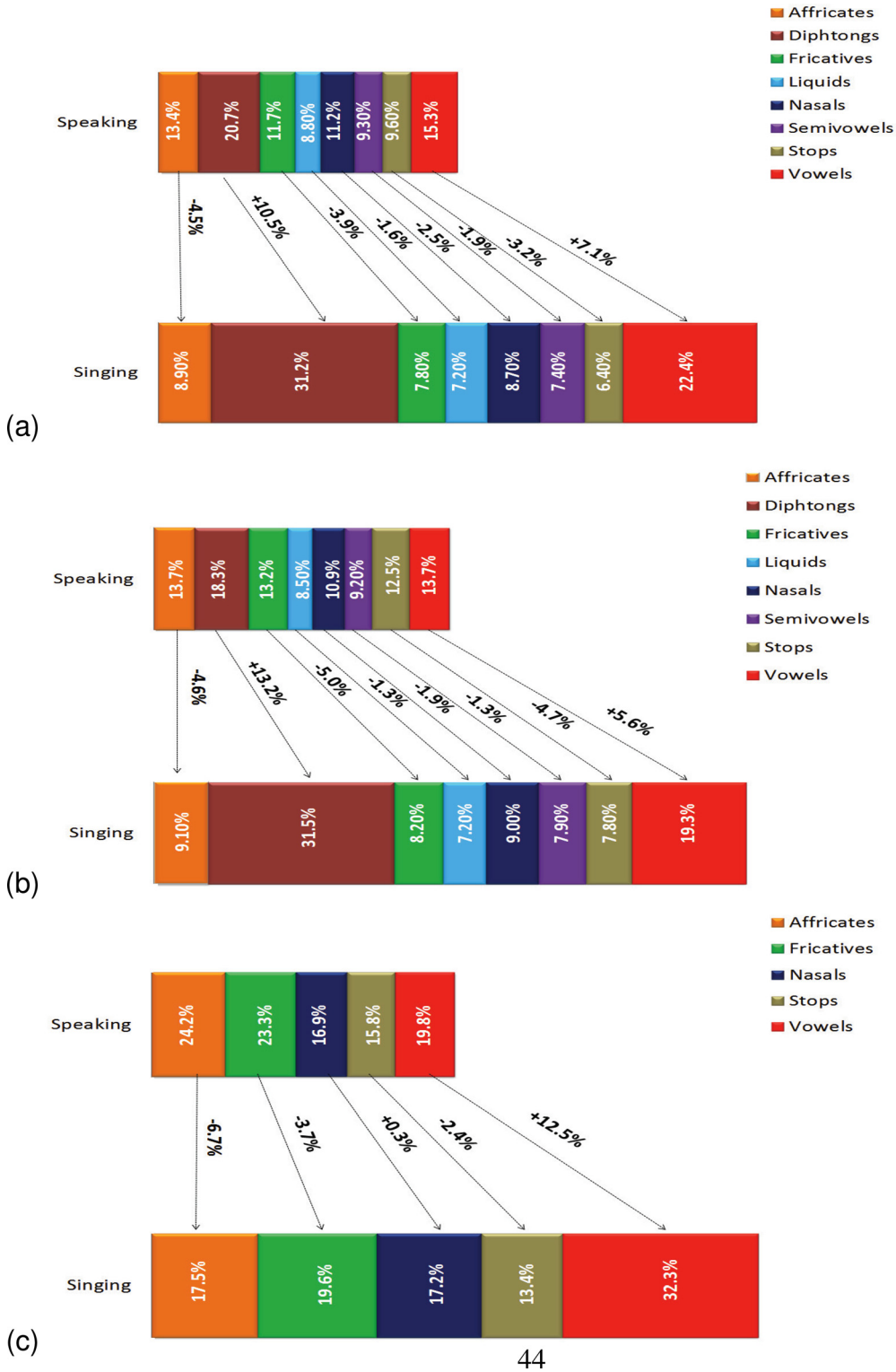


FIG. 2. (Color online) A comparison between phoneme class duration, assuming an ideal word containing a single entry from each phoneme class. Each phoneme class is noted by colors for that language; percentages show phone class duration for speaking and singing in each language. (a) English, (b) Hindi, and (c) Farsi.

singing voice conversion (e.g., given spoken lyrics, convert that stream to a singing audio stream). In New *et al.* (2010), the authors use different speech parameters, such as fundamental frequency, duration, and spectral features, to produce a singing voice from spoken speech. Knowing the differences in the duration of the phonemes is helpful in transforming duration features and improving the naturalness of the resulting generated singing voice.

B. Fundamental frequency

In neutral speech, fundamental frequency variations can reflect lexical stress, cognitive stress, or the emotional state

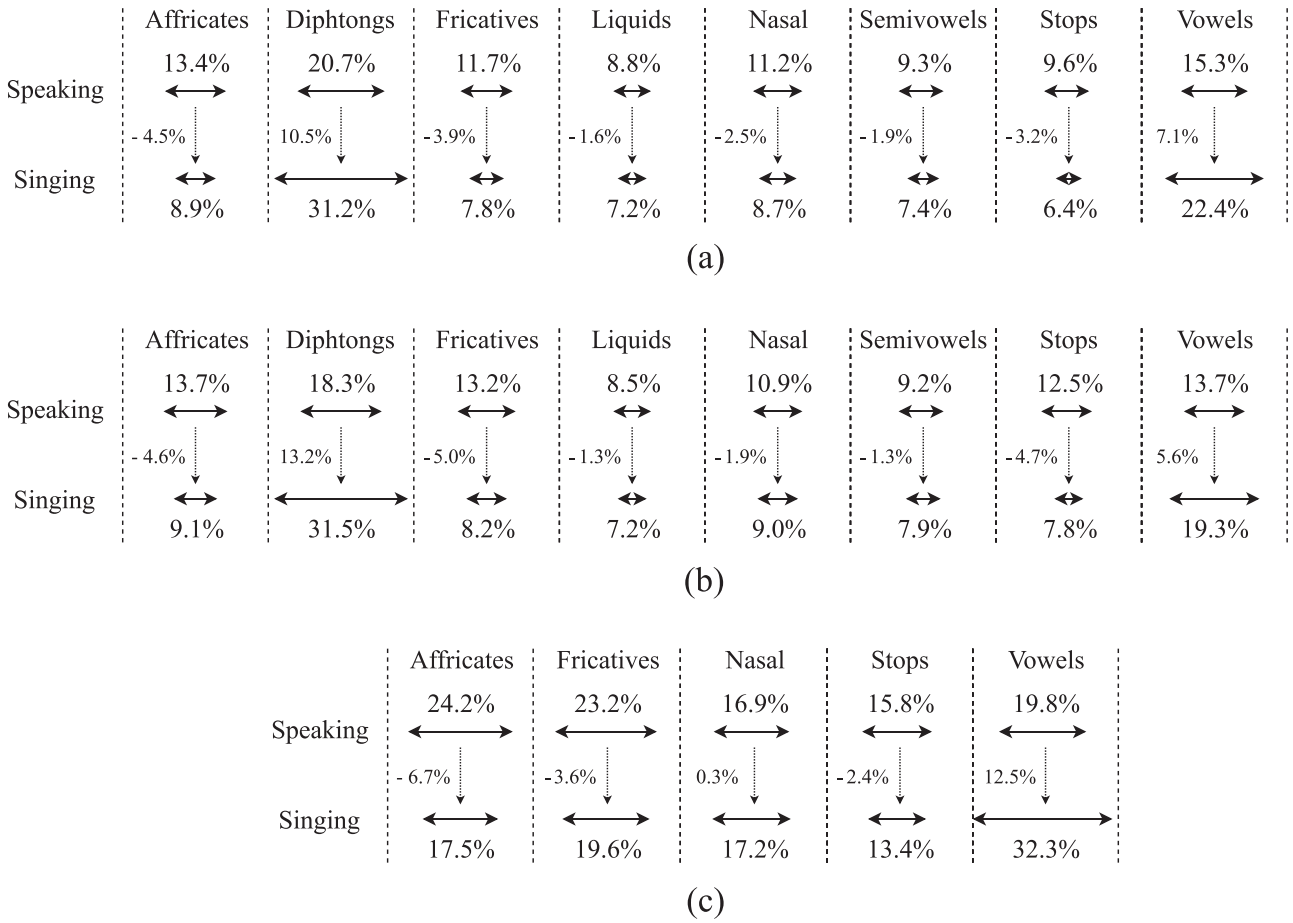


FIG. 3. An ideal word (one phoneme from each class) produced in both speaking and singing with their corresponding phoneme class duration coverage plus percent change from speak-to-sing. (a) English, (b) Hindi, and (c) Farsi.

of a speaker and add intelligibility to the spoken words. In singing, the fundamental frequency can be up to three octaves higher (Loscos *et al.*, 1999). One functionality of fundamental frequency in singing is to represent the pitch of a certain musical note to maintain the musical melody of a song. The study in New *et al.* (2010) considered mean fundamental frequency, overshoot, and vibrato to convert spoken text to singing voice. The mean fundamental frequency was subtracted from the *F*0 contour of a speaker to create pure *F*0-fluctuations. The study by Natke *et al.* (2003) also stated that an internal reference for pitch-matching exists in singing and although a similar version of this internal reference exists in spoken speech, it is not as strong. It is stated that while trained singers will match the pitch to a musical note almost perfectly even untrained singers will also try to match their *F*0 with a desired target singing *F*0. The failure to do so results in what is referred to as a "bad" glissando (Natke *et al.*, 2003).

Compared to speaking, singing has a higher fundamental frequency with a larger dynamic range (Loscos *et al.*, 1999). In our analysis, we were interested in observing the change in mean fundamental frequency that occurs from speaking to singing for vowels /a:/, /e:/, /i:/, /o:/, and /u:/. The selected vowels belong to approximately the same production space across each language. Moreover, we were also interested in comparing the variation of the fundamental frequency in different languages. For *F*0 extraction, we employed Wavesurfer—a speech analysis toolkit. We used an analysis window length of 200 samples (25 ms) with a skip frame rate of 100 samples (12.5 ms).

In addition to the mean fundamental frequency, we also calculated the coefficient of variation (CV) across six speakers in each language. The CV was included in order to show the dispersion of our variables. Figure 4 shows the results of this experiment. According to Fig. 4, the mean *F*0 of spoken vowels is significantly lower than that of the sung vowels for all three languages. The CV values obtained for the spoken vowels (excluding the English vowel /i:/) are also significantly lower than those obtained for the sung vowels. We also plotted cross-language results of speaking and singing in Fig. 5.

In general, we showed that Hindi speakers had the highest fundamental frequency in both singing and speaking, whereas English speakers had the lowest. This change could be due to the fact that the Hindi songs are usually sung in higher pitch compared to English and Farsi songs (Trehub *et al.*, 1993). This suggests that Hindi singers attempt to match a higher fundamental frequency compared to the other language speakers.

Our results also showed that the vowel /o:/ has the highest fundamental frequency during speaking in all

https://doi.org/10.1121/10.0001526

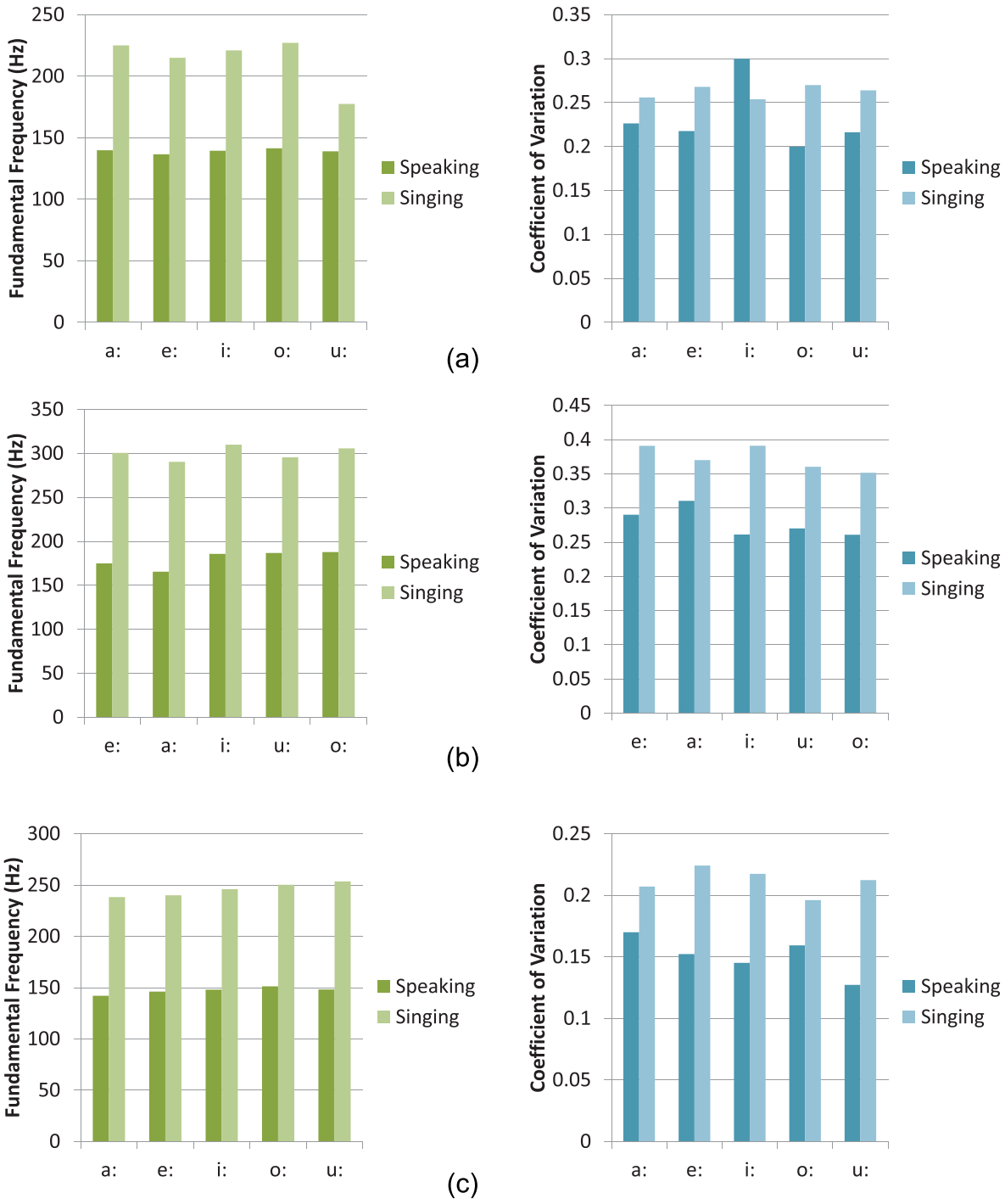


FIG. 4. (Color online) Fundamental frequency (*F*0) and coefficient of variation (CV) for speaking and singing the same text content (within each language) for (a) American English, (b) Hindi, and (c) Farsi speakers. (a) American English, speaking and singing; (b) Hindi, speaking and singing; and (c) Farsi, speaking and singing.

three languages. However, this result was not true for singing. Vowels /i:/, /o:/, and /u:/ showed the highest fundamental frequency in singing of Hindi, English, and Farsi, respectively. This result suggests that the speakers also attempt to match the different vowels during singing, which leads to more variations in the fundamental frequencies. However, in speaking, there was no predefined pitch and, therefore, the average pitch across different vowels in different languages shows less variations.

Our results in this domain are important in that they show how the fundamental frequency changes from speaking to singing in different languages. These results can provide information for voice conversion systems that rely on time duration and fundamental frequency changes in vowels from speaking to singing.

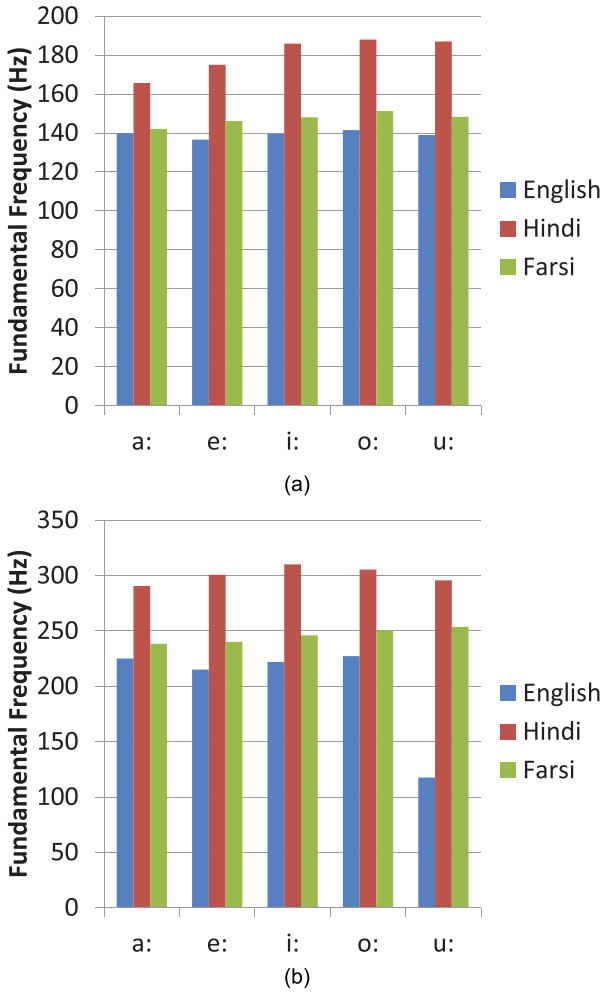


FIG. 5. (Color online) Cross-language fundamental frequency for (a) speaking and (b) singing for the same text content. (a) Speaking text content and (b) singing text content.

V. SPEAKER-DEPENDENT FORMANT ANALYSIS

Apart from the temporal changes, singing also introduces acoustic differences in the frequency structure of a speaker's voice (see Fig. 1). In the frequency domain, the greatest change happens in the formant frequencies of the vowels. Formant frequencies are well known as the spectral peaks of the speech spectrum, which represents the resonances of the vocal tract (Deller *et al.*, 2000; O'Shaughnessy, 2000). It has been shown that trained singers tend to tune their first two resonances of the vocal tract ($R1$ and $R2$) to a multiple of their fundamental frequency (Henrich *et al.*, 2011). It is suggested that this can enhance the overall sound level of the voice, which is necessary for professional singers. A study by Sundberg (1977) showed that moving the articulatory organs can significantly change the frequencies of the formants (specifically, the first two formants). Again, it is well known in speech modeling that each articulatory configuration corresponds to a set of formant frequencies, which, in turn, results in a particular vowel sound (Sundberg, 1977). In a study by Fox and Jacewicz (2008), the authors examined the vowel space areas of different dialects used in central Ohio, south-central Wisconsin, and western North Carolina. Given the large variation of the formant frequencies of different vowels, vowel space area is traditionally characterized by the area inside of a triangle formed by three corner vowels in the acoustic space ($F1$-$F2$ formant frequency plane; Deller *et al.*, 2000). This area is also characterized as the "working area" of the vowel system (Fox and Jacewicz, 2008). A small vowel space area shows that the formant frequencies of different vowels are close to each other and, therefore, vowels have low variations. In this study, we also compare the change in vowel space areas of speaking and untrained singing speakers.

A. Vowel space area

Inspired by Fox and Jacewicz (2008), we determined the vowel space area of three known vowels in English, Hindi, and Farsi. We selected three vowels (i.e., /a:/, /e:/, and /u:/) which belong to different articulatory production spaces in the three languages. We defined the vowel acoustic space by taking the average of the first and second formants of the vowels. We used Wavesurfer with a window length of 200 samples (25 ms) and a frame skip rate of 100 samples (12.5 ms) for formant extraction.

Vowel space area refers to the two-dimensional area in the $F1$-$F2$ plane, where $F1$ and $F2$ are the first two formants. Vowel space area is bounded by lines that connect the first and second formant frequency coordinates of vowels. A common way of calculating this area is to make static measurements of the $F1$ and $F2$ values for each of the four corner vowels (or three-point vowels, /a,i,u/ for a triangle) at the 50% vowel duration for multiple productions of each vowel. The mean $F1$ and $F2$ values for each of the four corner vowels are then used to compute the area of the quadrilateral formed by the corner vowels. The vowel space area

TABLE II. Speaker-dependent vowel space area of singing and speaking for English, Hindi, and Farsi.

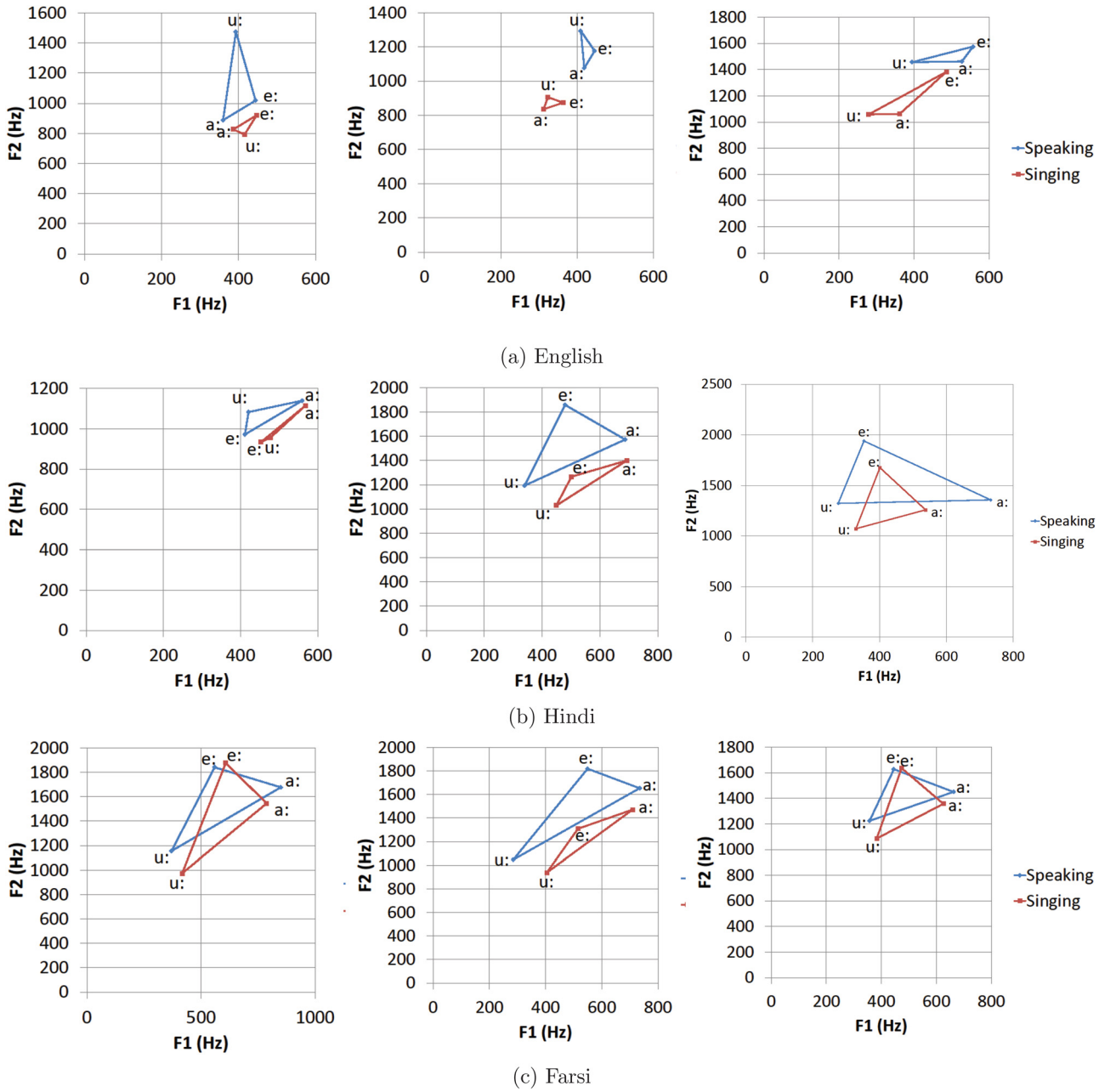
Language	Subjects	Speaking	Singing	Singing/speaking
English	Speaker 1	22 537.12	2404.05	0.10
	Speaker 2	6014.20	2357.23	0.39
	Speaker 3	3392.04	1592.76	0.49
	Speaker 4	7451.56	12 896.29	1.73
	Speaker 5	5385.39	1450.42	0.26
	Speaker 6	1226.18	5020.12	4.09
Hindi	Speaker 1	7463.29	929.81	0.12
	Speaker 2	14 643.35	1141.33	0.07
	Speaker 3	89 062.59	19 283.5	0.21
	Speaker 4	90 185.12	23 762.8	0.26
	Speaker 5	205 267.8	86 505.19	0.42
	Speaker 6	139 782.2	56 271.64	0.40
Farsi	Speaker 1	114 985.3	112 590.4	0.97
	Speaker 2	120 499	26 984	0.22
	Speaker 3	3328.55	46 164.05	13.86
	Speaker 4	68 913	87 203.08	1.26
	Speaker 5	51 541.54	54 283.84	1.05
	Speaker 6	17 133.26	48 472.97	2.82

https://doi.org/10.1121/10.0001526

is known to be an acoustic proxy for the kinematic displacements of the articulators. Sandoval (2013) explained the importance of this measure. They also introduced different ways for calculating this measure.

Table II summarizes the speaker-dependent vowel acoustic spaces for three female speakers and three male speakers in each language. The formant space ratio between singing and speaking is also shown in Table II. In Fig. 6, we visualize the same results for two female speakers and one male speaker in the $F1$-$F2$ plane.

The results from Fig. 6 show that all speakers (excluding one Farsi male speaker) have smaller vowel spaces during singing versus speaking. Table II also confirms this result. Although the formant space ratio of sing/speak is less than one for almost all speakers, dependencies between speakers are still observable. The first two English female speakers have a lower sing/speak formant space ratio than the English male speaker has. Except for one English male speaker, the rest of the male speakers show significantly higher sing/speak ratios than the English female speakers show, meaning that the formant space area of English male speakers in singing is larger than in speaking. This result introduces a gender dependency of the vowel spaces in the formant frequency plane, which is reasonable considering the difference in the formant frequencies between the two genders when speaking. Moreover, this dependency exists when we consider the male and female speakers of Hindi and Farsi as well. We can see from Table II that in both



(a) English

(b) Hindi

(c) Farsi

FIG. 6. (Color online) Vowel space areas for two females (two left figures) and one male (one right figure) in (a) English, (b) Hindi, and (c) Farsi.

Hindi and Farsi, the ratio of sing/speak is generally higher for the male speakers.

To summarize, we have shown that female speakers in all three languages had smaller vowel space areas in singing than in speaking. An exception was Speaker 9 in Table II, who shows a significantly larger vowel space area. In contrast, the male speakers showed a significantly larger vowel space area in singing than in speaking. This result applies to speakers of all three languages. Our results indicate that female speakers tend to produce vowels in a more compact vowel space area during singing. These vowels are closer together in the formant space and are less separable than in speaking. Male speakers tend to have larger vowel space areas during singing, which means that two or more of the vowels are more widely spread apart in frequency and, therefore, more easily separable and distinguishable. During speaking, the formant frequencies vary less from one speaker to another. However, during singing, this variability increases and is gender dependent.

B. Spectral shift

Next, we considered analysis of the spectral shift of vowels in the $F1$-$F2$ plane from speaking to singing. In order to calculate this shift, we estimated the $F1$-$F2$ distances between the vowels of speaking and the vowels of singing for each speaker. The averages of the distances across male and female speakers are summarized in Table III.

Spectral shift or the average distance between the first two formants in the vowels is an important measure that determines the shape of the spectral envelop. This measure varies significantly from different vowels; therefore, it is an effective feature for speech recognition. This measure also changes for different speakers, specifically for different genders; therefore, it can also be used for speaker verification/ID and gender recognition tasks. Here, we benefit from this measure to quantify the differences between the spectral features of speaking and singing.

As realized from Table III, the vowel /a:/ has the lowest spectral shift across both male and female speakers in all three languages. The largest spectral shift is in the vowel /u:/ for both English and Farsi speakers, whereas for Hindi, the vowel /e:/ has the largest shift. In English and Farsi, female speakers show a higher spectral shift than male speakers do for all vowels (except the vowel /a:/ for Farsi female speakers). However, for male speakers of the Hindi language, all vowels show a higher spectral shift than do those for the female speakers.

These results show that although male speakers tend to have large sing/speak vowel space ratios, the overall spectral shift for each vowel is, generally, lower than that for the females. This implies that untrained male singers tend to have more distinguishable vowels during singing, and the position of these vowels does not change significantly from the speaking production space compared to that for the female untrained singers. The exceptions in this case are the Hindi male singers, who tend to have a lower spectral shift than their Hindi female counterparts. The Hindi male singers also had considerably lower sing/speak formant frequency ratios than the English and Farsi males.

TABLE III. Spectral shift between singing and speaking of vowels /a:/, /e:/, and /u:/ for English, Hindi, and Farsi.

Language	Speech style	a:	e:	u:
English	Speaking	128.33	129.30	373.71
	Singing	215.06	57.52	227.43
Hindi	Speaking	98.93	294.52	169.95
	Singing	231.85	330.43	232.71
Farsi	Speaking	188.08	209.50	236.90
	Singing	57.74	84.50	107.57

VI. KULLBACK-LEIBLER DIVERGENCE (KLD) BETWEEN SINGING AND SPEAKING

In order to explore the degree of human production space similarity between singing and speaking, we trained Gaussian mixture models (GMMs) on both speech styles. We, then, used the KL divergence metric to quantitatively assess the difference between the two Gaussian acoustic style models. The KL divergence is used widely in areas such as pattern recognition and machine learning. It represents a composite distance metric between two distributions and, therefore, can be used to represent the similarity between two acoustic models (Hershey and Olsen, 2007). The KLD measure is another criterion for quantifying the differences between the spectral envelop of speaking versus singing voices.

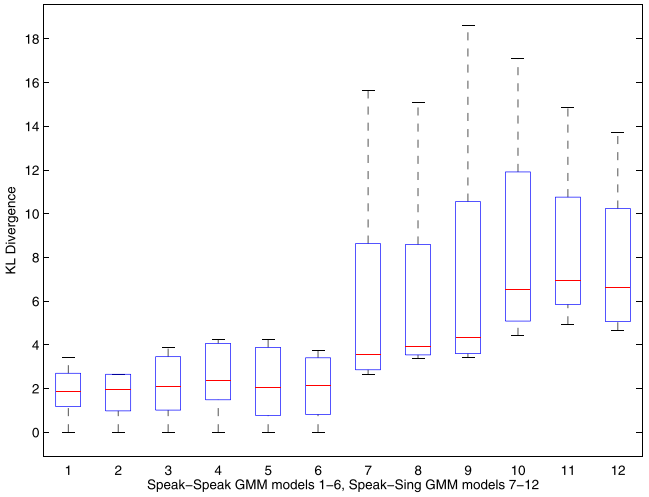
Calculating the KL divergence between two GMMs is not straightforward for multiple Gaussians in each model; Goldberger *et al.* (2003) proposed a Monte Carlo based approach to approximate the KL divergence between two GMMs. Their proposed method has been used in various applications. For example, Ramirez *et al.* leveraged the Monte Carlo based estimation of the KL divergence to calculate the distance between speech and noise distributions for a voice activity detector (VAD) system (Ramírez *et al.*, 2004); their results showed a significant improvement over the previous VAD algorithms.

In this study, we used an approximate value of the KL divergence, known as the "relative entropy." The KL divergence between two distributions, $f(x)$ and $g(x)$, is defined as

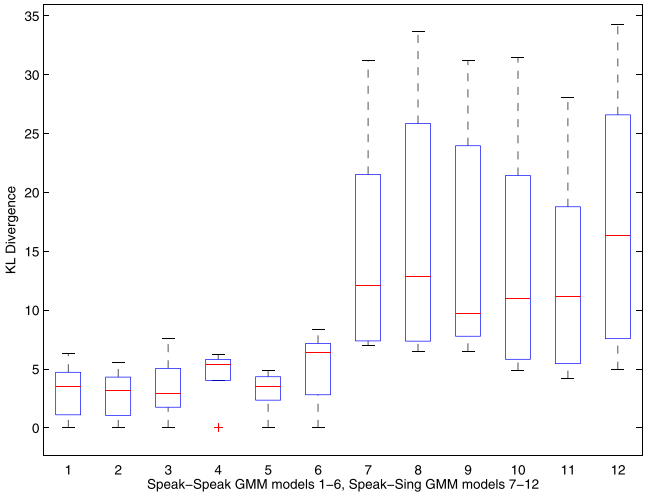
$$D(f \parallel g) = \int f(x)(\log(f(x)) - \log(g(x)))dx. \quad (1)$$

First, we trained a separate GMM model for each speaker in each language on neutral speech and singing. We used 36-dimensional MFCC features. We chose this feature set as it is the most widely used in speech recognition and SID applications (Reynolds, 1995; Deller *et al.*, 2000). We, then, used a KL divergence metric to find the distance between the GMM models trained for singing and speaking. We finally compared this result with the KL divergence of GMM models trained only on neutral speech. The results are plotted in Fig. 7. In Fig. 7, we show the box plot

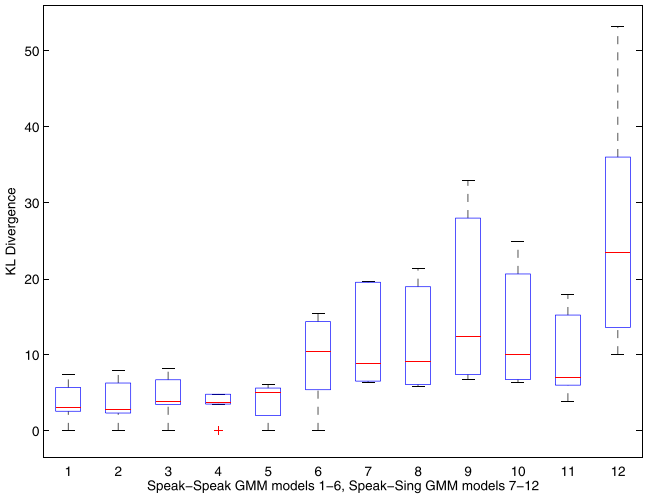
https://doi.org/10.1121/10.0001526



(a)



(b)



(c)

FIG. 7. (Color online) KL-divergence measures between cross-model GMMs and same model GMMs for (a) English, (b) Hindi, and (c) Farsi. Note: model-pair comparisons 1–6 are speak-to-speak GMM KL divergence distances; model-pair comparisons 7–12 are speak-to-sing KL divergence distances. (a) English, (b) Hindi, and (c) Farsi.

representations of all the KL divergence values for six speakers of each language separately. The mean KL divergence for a certain speaker is denoted by a red horizontal line on the plot; the rest of the box distribution shows the range of KL divergence values when a speaker was compared to other speakers of the same language (including themselves). Results 1–6 show the KL divergence of speaker-dependent GMM models trained on only neutral spoken speech, and results 7–12 show the KL divergence for the GMM models trained on neutral speech and singing.

As seen from Fig. 7, the mean KL divergence of each speaker increases significantly from speaking to singing, suggesting a meaningful change in spectral based speech production between the two speech styles (note that the text context of both styles is the same). The Hindi and Farsi languages produce high KL divergences in cross-model comparison between singing and speaking (mean KL divergences are 15.27 and 14.91 for Hindi and Farsi, respectively); however, this value is significantly lower for the English speakers (the mean KL divergence for English is 7.47). These results indicate that the spectral features of Hindi and Farsi speakers have greater mismatch production variations compared to the spectral features of English speakers.

We have also plotted these results in a three-dimensional (3-D) contour plot (see Fig. 8) in order to visually express the change in KL divergence when changing the speaking style from regular speech to singing and show how this change is reflected across the three languages. Figure 8 shows the drastic change from neutral speech to singing, and we can also see that this change is significantly greater for the Hindi and Farsi languages.

VII. SID OF SINGING

In the area of music retrieval, dealing with large music datasets requires building automated systems that perform classification based on factors such as musical content, lyrics, song genres, and singers. In the case of classifying musical content based on different singers, singer ID systems are needed. Such applications aim to recognize the singer of a song by analyzing the audio features of the music signal (Zhang, 2003). Because of the time-frequency differences between the speaker during singing versus the speaker during speaking, automatic SID systems degrade in performance when introduced with singing speech.

In this section, we illustrated an application of SID using an open set system and its performance when singing speech is applied. For this case, we used a GMM-UBM based system. We noted that other SID solutions exist, such as *i*-vector-PLDA, *X*-vector, or *t*-vector (Hansen and Hasan, 2015); however, GMM-UBM was selected because it allows for direct comparison of GMM models using the spectral structure. Therefore, the goal here was to quantify changes from neutral spoken speech in SID. A universal background model (UBM) was constructed using TIMIT data with 438 male speakers and 192 female speakers. We employed 1024 mixtures to train the UBM. Next, we obtained a maximum

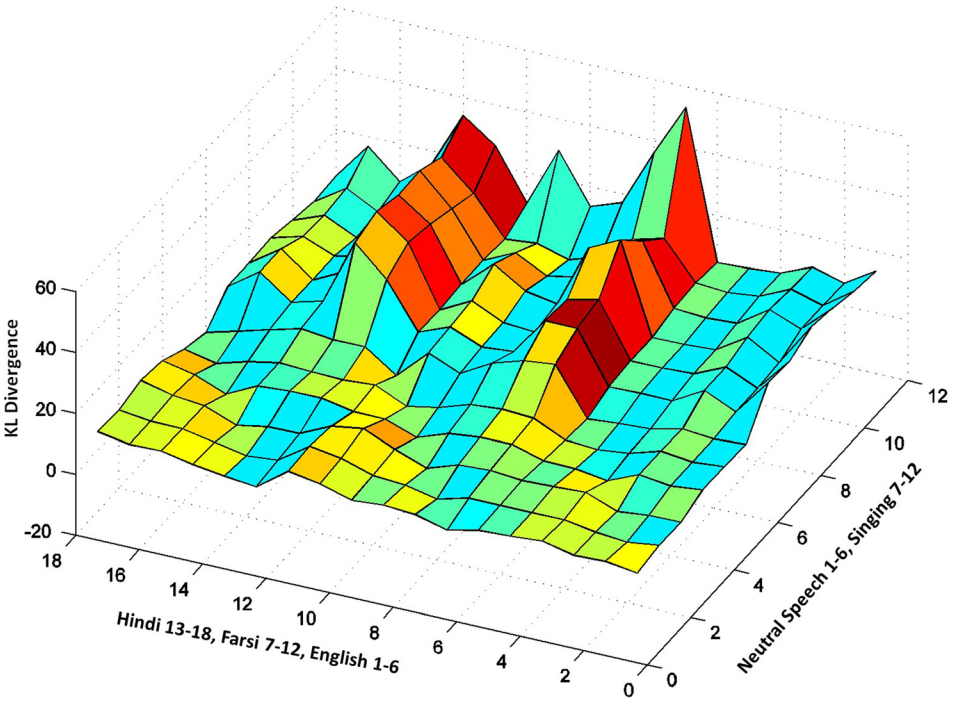


FIG. 8. (Color online) 3-D plot KL divergences between cross-model GMMs and same model GMMs for English, Farsi, and Hindi.

a posteriori adapted Gaussian mixture model for each of our English, Hindi, and Farsi speakers. For maximum *a posteriori* adaptation, we used 17 speakers of English, 17 speakers of Hindi, and 13 speakers of Farsi.

A. Speech processing and features

All speech data were down-sampled from the original sampling rate of 44 kHz to 8 kHz. A speech activity detection (SAD) system was applied in order to remove all silence frames. We used 36-dimensional MFCC features (12 static, 12 delta, and 12 delta-delta). MFCCs are widely used features in speech recognition and SID systems because of their ability to effectively represent the characteristics of the vocal tract (Deller *et al.*, 2000; Reynolds, 1995).

B. Training and testing

In order to consider the implications of singing on SID, we trained three SID systems for each language separately. Since the acoustic phoneme space for each language differs, the resulting models would potentially require slightly different settings. For training, we used 22 tokens, and for testing, we used 11 tokens per speaker. The duration of each token was 10 s long. We hypothesize that the identity of a speaker can be determined using a 10-s speech segment (see Hansen *et al.*, 2013; Angkititrakul and Hansen, 2007; Suh and Hansen, 2012; Prakash and Hansen, 2007, for studies on short duration speaker ID). We split our dataset into train and test by allocating 2/3 and 1/3 of the data for training and testing, respectively. We obtained 9500 test trials for each of the three languages. Test trials represent a subset of the dataset used to evaluate the performance of the speaker recognition system.

C. Results

Table IV summarizes the results of the SID experiment. We used the equal error rate (EER) to quantify the accuracy of the systems. According to Table IV, we obtained better performance when we train and test all three systems (English, Hindi, and Farsi) with neutral speech data. The system performance degraded significantly when we tested our system on singing recordings.

We also observed that the Hindi SID system performed significantly better than the English system and slightly better than the Farsi system when trained and tested with neutral speech. However, Hindi showed the highest degradation in performance (an increase in EER of +30.54%) when we introduce singing test data. Farsi ranked second with an EER increase from 12.85% to 36.67%, a +23.81% absolute loss in performance, and English has the lowest degradation with an EER change of +20.17%. This result agrees with our previous results on acoustic model variation based or KL divergence, indicating that Hindi and Farsi models showed higher dissimilarity between singing and speaking.

VIII. CONCLUSION

Speech production variability plays a significant role in the effectiveness of models for speech technology, so analysis of, modeling, and quantifying these changes offer

TABLE IV. The equal error rate (EER) of the SID system trained on neutral speech and singing for English, Farsi, and Hindi speakers.

Train	Test	English	Farsi	Hindi
Read	Read	18.18	12.85	11.84
Read	Sing	38.35	36.67	42.39

opportunities to develop more effective compensation methods for sustained system performance (Hansen, 1996; Hansen and Hasan, 2015). This study focused on one speech production mismatch domain of speaking versus singing the same text content. Here, we studied changes in speech production between speaking and untrained karaoke singing within the context of three languages (English, Hindi, and Farsi). The majority of previous studies on singing considered only trained singers (Watts *et al.*, 2006; Bloothooft and Plomp, 1986; Carlsson and Sundberg, 1992) with various goals, including assessing the quality of trained singers. However, singing involves an alternate neural processing pathway versus speaking and it is, therefore, of scientific interest to understand speaking/singing differences. Also, singing is a process that occurs in a wider range of contexts than just music entertainment and for these cases, there is a need for assessment and analysis of untrained singing.

In this study, we analyzed the temporal changes that occur across phonetic classes when we alternate from neutral speech to singing using the same text content. Our results indicate that the biggest increase in duration occurs for vowels and diphthongs across all three languages. In this context, we studied fundamental frequency changes and vowel space areas in the $F1$-$F2$ (first two formants) plane for different speakers of English, Hindi, and Farsi.

As a result of our experiments, we observed that English, Farsi, and Hindi speakers attempt to match the vowels /o:/, /u:/, and /i:/ at a higher pitch, respectively. These vowels have the highest fundamental frequency during singing. Our prosodic analysis is important because it can be used in speech to singing conversion applications where speech parameters, such as fundamental frequency, duration, and spectral coefficients, are transformed to generate a singing voice from a spoken speech signal (New *et al.*, 2010). We also observed that mean fundamental frequencies are higher for singing compared to normal speaking, and this difference is more significant for Hindi speakers compared to speakers of English and Farsi. We hypothesize this result is mainly because the Hindi songs usually contain higher pitch signals, which suggests that Hindi speakers attempt to match a higher fundamental frequency while singing.

We also showed that for female speakers, the vowel space area of singing is, generally, smaller than that of speaking, which means that vowels in singing are positioned closer to each other in the $F1$-$F2$ formant space and as a result are harder to differentiate. Males usually tend to have a greater vowel space area in singing than in speaking in all three languages. Finally, we used the KL divergence metric to assess changes in the speech production space, representing dissimilarities between singing and speaking for English, Hindi, and Farsi. Our results indicated that Hindi and Farsi have the highest dissimilarity between singing and speaking. If we compare these results with our SID results, we can see that the accuracy of SID solutions will degrade significantly between spoken versus singing and this change is more pronounced in the Hindi and Farsi languages than it is in English.

This paper explains some of the differences between acoustic characteristics of speaking and singing voices. Introducing these acoustic differences is the first step toward improving speech processing applications for untrained singing sounds. Further studies and experiments are required to improve speech processing applications. For example, this paper shows that there is a significant difference between spectral features of singing and speaking. Therefore, one way to improve speech processing applications is to first identify if the signal is untrained singing and, then, convert its spectral features in a similar way to the speaking style. The results of this study and similar studies show the differences between spectral features and, therefore, can be useful in developing the conversion system.

ACKNOWLEDGMENTS

This work was supported by the Grant No. R01 DC016839-03 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

Angkititrakul, P., and Hansen, J. H. L. (**2007**). "Discriminative in-set/out-of-set speaker recognition," IEEE Trans. Audio, Speech, Lang. Process. **15**(2), 498–508.

Barrichelo, V. M. O., Heuer, R. J., Dean, C. M., and Sataloff, R. T. (**2001**). "Comparison of singer's formant, speaker's ring, and LTA spectrum among classical singers and untrained normal speakers," J. Voice **15**(3), 344–350.

Beeman, S. A. (**2017**). "Perceptions of voice teachers regarding students' vocal behaviors during singing and speaking," J. Voice **31**(1), 111.e19–111.e28.

Bloothooft, G., and Plomp, R. (**1984**). "Spectral analysis of sung vowels. I. variation due to differences between vowels, singers, and modes of singing," J. Acoust. Soc. Am. **75**(4), 1259–1264.

Bloothooft, G., and Plomp, R. (**1986**). "The sound level of the singer's formant in professional singing," J. Acoust. Soc. Am. **79**(6), 2028–2033.

Bou-Ghazale, S., and Hansen, J. H. L. (**2000**). "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Trans. Audio, Speech, Lang. Process. **8**(4), 429–442.

Brown, W. S., Rothman, H. B., and Sapienza, C. M. (**2000**). "Perceptual and acoustic study of professionally trained versus untrained voices," J. Voice **14**(3), 301–309.

Cairns, D., and Hansen, J. H. L. (**1994**). "Nonlinear analysis and detection of speech under stressed conditions," J. Acoust. Soc. Am. **96**(6), 3392–3400.

Carlsson, G., and Sundberg, J. (**1992**). "Formant frequency tuning in singing," J. Voice **6**(3), 256–260.

Cleveland, T. F., and Sundberg J. (**1999**). "Formant frequencies in country singers' speech and singing," J. Voice **13**(2), 161–167.

Deller, J. R., Hansen, J. H. L., and Proakis, J. G. (**2000**). *Discrete-Time Processing of Speech Signals* (IEEE, New York).

Fant, G. (**2012**). *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations* (de Gruyter Mouton, Berlin), Vol. 2.

Fowler, C. A., and Brown, J. M. (**1997**). "Intrinsic *f*0 differences in spoken and sung vowels and their perception by listeners," Percept. Psychophys. **59**(5), 729–738.

Fox, R. A., and Jacewicz, E. (**2008**). "Analysis of total vowel space areas in three regional dialects of American English," J. Acoust. Soc. Am. **123**(5), 3068.

Ghaffarzadegan, S., Boril, H., and Hansen, J. H. L. (**2016**). "Generative modeling of pseudo-whisper for robust whispered speech recognition," IEEE Trans. Audio, Speech, Lang. Process. **24**(10), 1705–1720.

Godin, K., and Hansen, J. H. L. (**2011**). "The effects of physical task stress on phone classes of American English," J. Acoust. Soc. Am. **130**(6), 3992–3998.

Goldberger J., Gordon S., and Greenspan, H. (**2003**). "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 487–493.
Hansen, J. H. L. (**1996**). "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," Speech Commun. **20**, 151–170.
Hansen, J. H. L., and Bou-Ghazale, S. (**1995**). "Duration and spectral based stress token generation for keyword recognition using hidden Markov models," IEEE Trans. Audio, Speech, Lang. Process. **3**(5), 415–421.
Hansen, J. H. L., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: A tutorial review," IEEE Signal Process. Mag. **32**(6), 74–99.
Hansen, J. H. L., Kim, W., Rahurkar, M., Ruzanski, E., and Meyerhoff, J. (**2011**). "Robust emotional stressed speech detection using weighted frequency subbands," EURASIP J. Adv. Signal Process. **2011**, 906789.
Hansen, J. H. L., Ruzanski, E., Boril, H., and Meyerhoff, J. (**2012**). "TEO-based speaker stress assessment using hybrid classification and tracking schemes," Int. J. Speech Technol. **15**(3), 295–311.
Hansen, J. H. L., Suh, J.-W., and Leonard, M. R. (**2013**). "In-set/out-of-set speaker recognition in sustained acoustic scenarios using sparse data," Speech Commun. **55**(6), 769–781.
Hansen, J. H. L., and Varadarajan, V. (**2009**). "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," IEEE Trans. Audio, Speech, Lang. Process. **17**(2), 366–378.
Henrich, N., Smith, J., and Wolfe, J. (**2011**). "Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones," J. Acoust. Soc. Am. **129**(2), 1024–1035.
Hershey, J. R., and Olsen, P. A. (**2007**). "Approximating the Kullback-Leibler divergence between Gaussian mixture models," in *IEEE-2007: Inter. Conf. on Acoustics, Speech and Signal Processing*, Vol. 4, pp. iv–317.
Khorram, S., Gideon, J., McInnis, M. G., and Provost, E. M. (**2016**). "Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge," in *ISCA INTERSPEECH-2016*, pp. 1215–1219.
Khorram, S., Jaiswal, M., Gideon, J., McInnis, M., and Provost, E. M. (**2018**). "The priori emotion dataset: Linking mood to emotion detected in-the-wild," in *ISCA INTERSPEECH-2018*, pp. 1903–1907.
Khorram, S., Sameti, H., Bahmaninezhad, F., King, S., and Drugman, T. (**2014**). "Context-dependent acoustic modeling based on hidden maximum entropy model for statistical parametric speech synthesis," EURASIP J. Audio, Speech, Music Process. **2014**(1), 12.
Khorram, S., Sameti, H., and King, S. (**2015**). "Soft context clustering for *f*0 modeling in hmm-based speech synthesis," EURASIP J. Adv. Signal Process. **2015**(1), 2.
Loscos, A., Cano, P., and Bonada, J. (**1999**). "Low-delay singing voice alignment to text," in *Proceedings of the ICMC*, Vol. 18.
Mehrabani, M., and Hansen, J. H. L. (**2011**). "Language identification for singing," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4408–4411.
Mehrabani, M., and Hansen, J. H. L. (**2013a**). "Dimensionality analysis of singing speech based on locality preserving projections," in *INTERSPEECH*, pp. 2910–2914.
Mehrabani, M., and Hansen, J. H. L. (**2013b**). "Singing speaker clustering based on subspace learning in the GMM mean supervector space," Speech Commun. **55**(5), 653–666.
Natke, U., Donath, T. M., and Kalveram, K. (**2003**). "Control of voice fundamental frequency in speaking versus singing," J. Acoust. Soc. Am. **113**(3), 1587–1593.
New, T. L., Dong, M., Chan, P., Wang, X., Ma, B., and Li, H. (**2010**). "Voice conversion: From spoken vowels to singing vowels," in *2010 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1421–1426.
O'Shaughnessy, D. (**2000**). *Speech Communication: Human and Machine* (IEEE Press, New York).
Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Švec, J. G., and Hillman, R. "Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function," Am. J. Speech Lang. Pathol. **27**(3), 887–905 (**2018**).
Patil, S., and Hansen, J. H. L. (**2010**). "The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification," Speech Commun. **52**, 327–340.
Peynircioğlu, Z. F., Rabinovitz, B. E., and Repice, J. (**2017**). "Matching speaking to singing voices and the influence of content," J. Voice **31**(2), 256.e13–256.e17.
Prakash, V., and Hansen, J. H. L. "In-set/out-of-set speaker recognition under sparse enrollment," IEEE Trans. Audio, Speech, Lang. Process. **15**(7), 2044–2052 (**2007**).
Ramírez, J., Segura, J. C., Benítez, C., de la Torre, A., and Rubio, A. J. (**2004**). "A new Kullback-Leibler VAD for speech recognition in noise," IEEE Signal Process. Lett. **11**(2), 266–269.
Reynolds, D. A. (**1995**). "Speaker identification and verification using Gaussian mixture speaker models,' Speech Commun. **17**(1-2), 91–108.
Saitou, T., Goto, M., Unoki, M., and Akagi, M. (**2007**). "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Vol. 15(7), pp. 2044–2052.
Sandoval, S. (**2013**). "Automatic assessment of vowel space area," J. Acoust. Soc. Am. **134**(5), 477–483.
Sangwan, A., Mehrabani, M., and Hansen, J. H. L. (**2011**). "Language identification using a combined articulatory prosody framework," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4400–4403.
Sjölander, K., and Beskow, J. (**2000**). "Wavesurfer—An open source speech tool," in *INTERSPEECH*, pp. 464–467.
Story, B. H. (**2004**). "Vowel acoustics for speaking and singing," Acta Acust. Acust. **90**(4), 629–640.
Suh, J.-W., and Hansen, J. H. L. (**2012**). "Acoustic hole filling for sparse enrollment data using a cohort universal corpus for speaker recognition," J. Acoust. Soc. Am. **131**(2), 1515–1528.
Sundberg, J. (**1974**). "Articulatory interpretation of the 'singing formant,'" J. Acoust. Soc. Am. **55**(4), 838–844.
Sundberg, J. (**1977**). "The acoustics of the singing voice," Sci. Am. **236**(3), 82–91.
Svec, J., and Granqvist, S. (**2010**). "Guidelines for selecting microphones for human voice production research," Am. J. Speech Lang. Pathol. **19**, 356–369.
Trehub, S. E., Unyk, A. M., and Trainor, L. J. (**1993**). "Maternal singing in cross-cultural perspective," Infant Behav. Dev. **16**(3), 285–295.
Tsai, W.-H., and Lee, H.-C. (**2012**). "Automatic singer identification based on speech-derived models," Proc. Int. J. Future Comput. Commun **1**(2), 94–96.
Tsai, W.-H., and Wang, H.-M. (**2004**). "Towards automatic identification of singing language in popular music recordings," in *Proc. 5th ISMIR*, pp. 568–576.
Tsai, W.-H., and Wang, H.-M. (**2007**). "Automatic identification of the sung language in popular music recordings," J. New Music Res. **36**(2), 105–114.
Watts, C., Barnes-Burroughs, K., Estis, J., and Blanton, D. (**2006**). "The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers," J. Voice **20**(1), 82–88.
Womack, B. D., and Hansen, J. H. L. (**1999**). "*N*-channel hidden Markov models for combined stress speech classification and recognition," IEEE Trans. Audio, Speech, Lang. Process. **7**(6), 668–677.
Zhang, T. (**2003**). "Automatic singer identification," in *Proceedings of the 2003 International Conference on Multimedia and Expo (ICME'03)*, IEEE, Vol. 1, pp. i–33.
Zhou, G., Hansen, J. H. L., and Kaiser, J. F. (**2001**). "Nonlinear feature based classification of speech under stress," IEEE Trans. Audio, Speech, Lang. Process. **9**(2), 201–216.